

FIGURE 1

Minimum signal peptide score	false positive rate	false negative rate	proba(0.1)	proba(0.2)
3,5	0,121	0,036	0,467	0,664
4	0,096	0,06	0,519	0,708
4,5	0,078	0,079	0,565	0,745
5	0,062	0,098	0,616	0,782
5,5	0,05	0,127	0,659	0,813
6	0,04	0,163	0,694	0,836
6,5	0,033	0,202	0,725	0,855
7	0,025	0,248	0,763	0,878
7,5	0,021	0,304	0,78	0,889
8	0,015	0,368	0,816	0,909
8,5	0,012	0,418	0,836	0,92
9	0,009	0,512	0,856	0,93
9,5	0,007	0,681	0,863	0,934
10	0,006	0,679	0,835	0,919

FIGURE 2

## Score curves

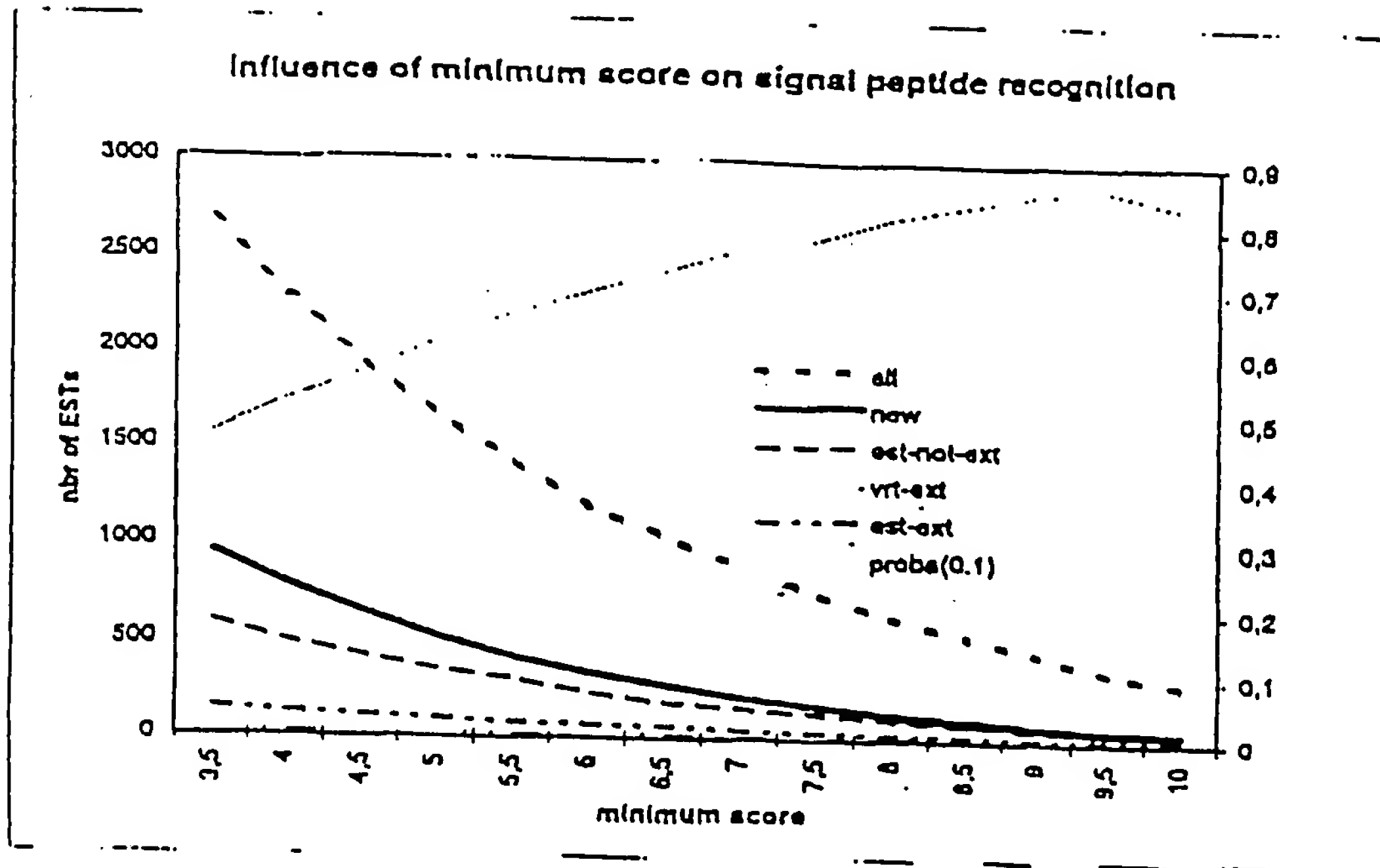


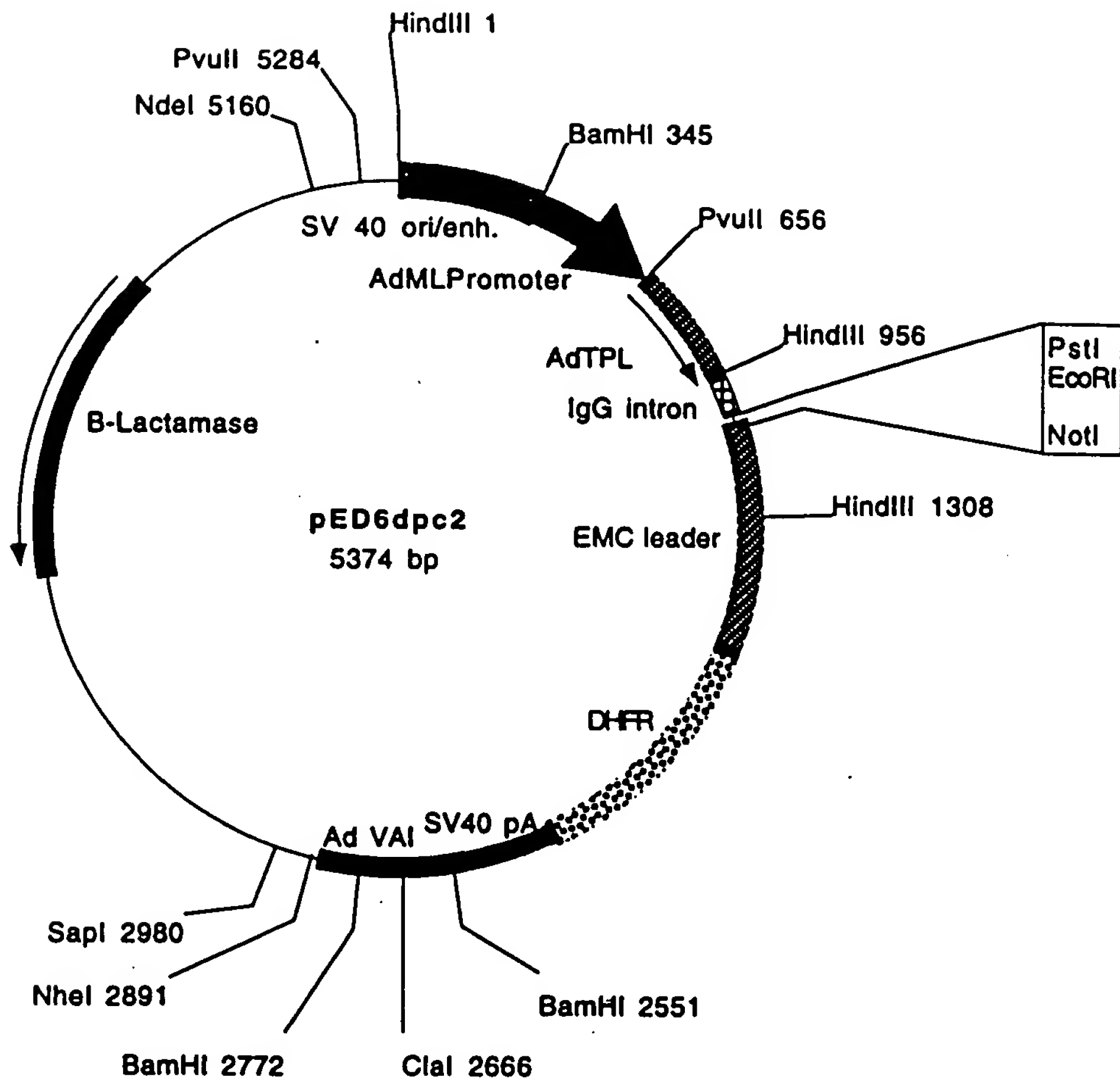
FIGURE 3

Minimum signal peptide score	All ESTs	New ESTs	ESTs matching public EST closer than 40 bp from beginning	ESTs extending known mRNA more than 40 bp	ESTs extending public EST more than 40 bp
3,5	2874	947	599	23	150
4	2278	784	488	23	126
4,5	1943	647	425	22	112
5	1657	523	353	21	96
5,5	1417	418	307	19	80
6	1180	340	238	18	68
6,5	1035	280	186	18	60
7	893	219	161	15	48
7,5	753	173	132	12	36
8	636	133	101	11	29
8,5	543	104	83	8	26
9	456	81	63	6	24
9,5	364	57	48	6	18
10	303	47	35	6	15

FIGURE 4

Tissue	ESTs		ESTs matching public EST closer than 40 bp from beginning	ESTs extending known mRNA more than 40 bp	ESTs extending public EST more than 40 bp
	All ESTs	New ESTs			
Brain	329	131	75	3	24
Cancerous prostate	134	40	37	1	6
Cerebellum	17	9	1	0	6
Colon	21	11	4	0	0
Dystrophic muscle	41	18	8	0	1
Fetal brain	70	37	18	0	1
Fetal kidney	227	116	48	1	19
Fetal liver	13	7	2	0	0
Heart	30	15	7	0	1
Hypertrophic prostate	86	23	22	2	2
Kidney	10	7	3	0	0
Large intestine	21	8	4	0	1
Liver	23	9	6	0	0
Lung	24	12	4	0	1
Lung (cells)	57	38	6	0	4
Lymph ganglia	163	60	23	2	12
Lymphocytes	23	6	4	0	2
Muscle	33	16	6	0	4
Normal prostate	181	61	45	7	11
Ovary	90	57	12	1	2
Pancreas	48	11	6	0	1
Placenta	24	5	1	0	0
Prostate	34	16	4	0	2
Spleen	56	28	10	0	1
Substantia nigra	108	47	27	1	6
Suprenals	15	3	3	1	0
Testis	131	68	25	1	8
Thyroid	17	8	2	0	2
Umbilical cord	65	17	12	1	3
Uterus	28	15	3	0	2
Non tissue-specific	568	48	177	2	28
Total	2677	947	601	23	150

FIGURE 5



Plasmid name: pED6dpc2

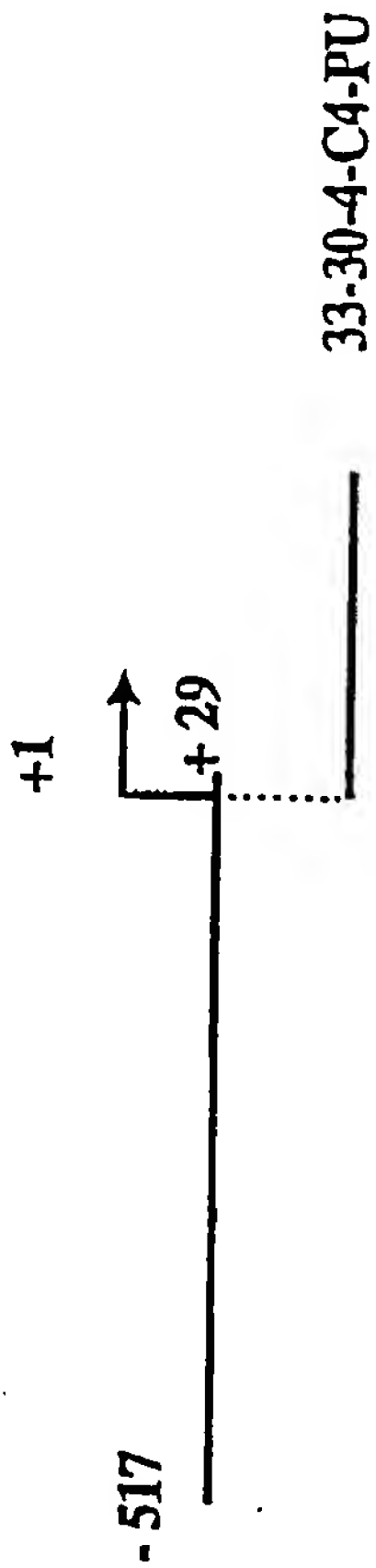
Plasmid size: 5374 bp

Comments/References: pED6dpc2 is derived from pED6dpc1 by insertion of a new polylinker to facilitate cDNA cloning. SST cDNAs are cloned between EcoRI and NotI. pED vectors are described in Kaufman et al.(1991), NAR 19: 4485-4490.

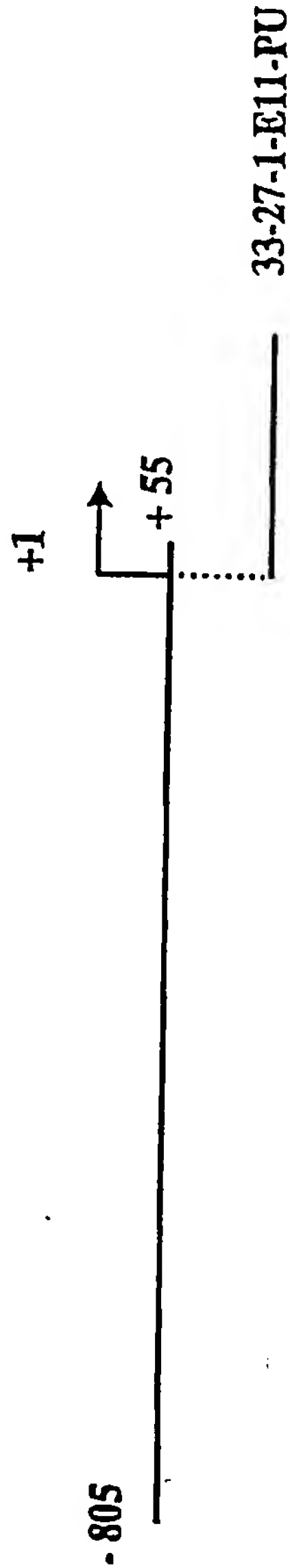
FIGURE 6

Description of Promoter structure isolated from SignalTag 5'ESTs

Promoter P13H2



Promoter P15B4



Promoter P29B6

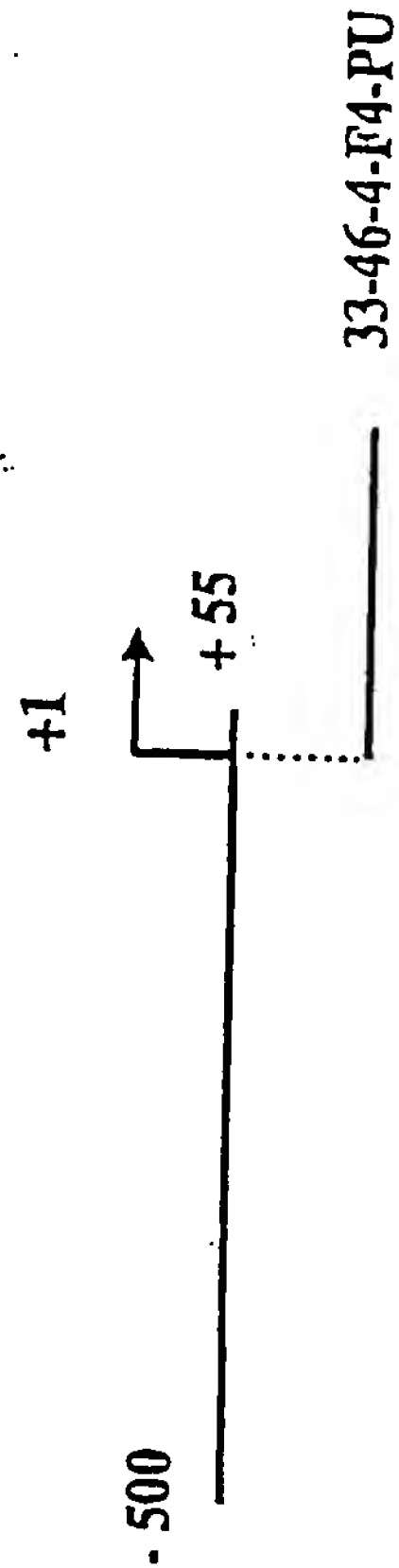


FIGURE 7

Description of Transcription Factor Binding Sites present on promoters isolated from SignalTag sequences.

Promoter sequence P13H2 (548 bp):

Matrix	Position	Orientation	Score	Length	Sequence
CMYB_01	-502	+	0.983	9	TGTCAGTTG
MYOD_Q8	-501	-	0.981	10	CCCAACTGAC
88_01	-444	-	0.960	11	AATAGAATTAG
88_01	-425	+	0.968	11	AACTAAATTAG
DELTAEF1_01	-390	-	0.980	11	GCACACOTCAG
GATA_C	-384	-	0.964	11	AGATAAATCCA
CMYB_01	-349	+	0.958	9	CTTCAGTTG
GATA1_02	-343	+	0.959	14	TTGTAGATAGGACA
GATA_C	-339	+	0.953	11	AGATAGGACAT
TAL1ALPHA47_01	-235	+	0.973	18	CATAACAGATGGTAAG
TAL1BETA47_01	-235	+	0.983	18	CATAACAGATGGTAAG
TAL1BETA1F2_01	-235	+	0.978	18	CATAACAGATGGTAAG
MYOD_Q8	-232	-	0.954	10	ACCATCTGTT
GATA1_04	-217	-	0.953	13	TCAAGATAAAGTA
IK1_01	-126	+	0.963	18	AGTTGGGAATTCC
IK2_01	-126	+	0.985	12	AGTTGGGAATTC
CREL_01	-123	+	0.982	10	TGGGAATTCC
GATA1_02	-98	+	0.960	14	TCAGTGATATGGCA
SRV_02	-41	-	0.951	12	TAAACAAAACA
E2F_02	-33	+	0.957	8	TTAGCGC
MZF1_01	-6	-	0.976	8	TGAGGGGA

Promoter sequence P16B4 (881 bp):

Matrix	Position	Orientation	Score	Length	Sequence
NFY_Q8	-748	-	0.958	11	GGACCAATCAT
MZF1_01	-738	+	0.962	8	CCTGGGGA
CMYB_01	-684	+	0.994	9	TGACCGTTG
VMYB_02	-682	-	0.985	9	TCCAACGGT
STAT_01	-673	+	0.988	9	TTCTTGGA
STAT_01	-673	-	0.951	9	TTCCAGGAA
MZF1_01	-658	-	0.958	8	TTGGGGGA
IK2_01	-451	+	0.965	12	GAATGGGAATTC
MZF1_01	-424	+	0.988	8	AGAGGGGA
SRV_02	-388	-	0.955	12	GAAAACAAAACA
MZF1_01	-216	+	0.960	8	GAAGGGGA
MYOD_Q8	-190	+	0.981	10	AGCATCTGCC
DELTAEF1_01	-178	+	0.958	11	TCCACCTTCC
88_01	6	-	0.982	11	GAGGCAATTAT
MZF1_01	16	-	0.988	8	AGAGGGGA

Promoter sequence P29B8 (555 bp):

Matrix	Position	Orientation	Score	Length	Sequence
ARNT_01	-311	+	0.984	16	GGA CTCACGTGCTGCT
NMYC_01	-309	+	0.985	12	ACTCACGTGCTG
USF_01	-309	+	0.985	12	ACTCACGTGCTG
USF_01	-309	-	0.985	12	OAGCAOQTGAGT
NMYC_01	-309	-	0.958	12	OAGCAOQTGAGT
MYC MAX_02	-309	-	0.972	12	CAGCACGTGAGT
USF_C	-307	+	0.997	8	TCACGTGC
USF_C	-307	-	0.991	8	GCACGTGA
MZF1_01	-282	-	0.988	8	CATGGGGA
ELK1_02	-105	+	0.983	14	CTCTCCGGAAGCCT
CET81P64_01	-102	+	0.974	10	TCCGGAAGCC
AP1_Q4	-42	-	0.963	11	AGTGA CTGAAC
AP1FJ_Q2	-42	-	0.981	11	AGTGA CTGAAC
PADS_C	45	+	1.000	9	TGTGGTCTC

Figure 8



```

10      20      30      40      50      60
SeqID214 MVIRVYIASSSGSTAIAKKKQDVLGFLEANKIGFEEKDIAANEENRKWMRENVPENSRPA
          ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
AF042081 MVIRVYIASSSGSTAIAKKKQDVLGFLEANKIGFEEKDIAANEENRKWMRENVPENSRPA
          10      20      30      40      50      60

70      80      90      100     110
SeqID214 TGNPLPPQIFNESQYRGDYDAFFEARENNAVYAFLGLTAPSGSKEAEVQAKQQ
          ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
AF042081 TGYPLPPQIFNESQYRGDYDAFFEARENNAVYAFLGLTAPPGSKEAEVQAKQQ
          70      80      90      100     110

```

FIGURE 9

seqID215	MADDLKRFLYKKLPSVEGLHAIVVSDRDGVPVIKVANDNAPEHALRPGFLSTFALATDQG
seqID185	MADDLKRFLYKKLPSVEGLHAIVVSDRDGVPVVKVANDNAPEHALRPGFLSTFALATDQG
AF082526	MADDLKRFLYKKLPSVEGLHAIVVSDRDGVPVIKVANDSAPEHALRPGFLSTFALATDQG
	*****
seqID215	SKLGLSKNKSIIICYNTYQVVQFNRLPLVVSFIASSSANTGLIVSLEKELAPLFEELRQV
seqID185	SKLGLSKNKSIIICYNTYQVVQFNRLPLVVSFIASSSANTGLIVSLEKELAPLFEELRQV
AF082526	SKLGLSKNKSIIICYNTYQVVQFNRLPLVVSFIASSSANTGLIVSLEKELAPLFEELIKV
	*****
seqID215	VEVS
seqID185	VEVS
AF082526	VEVS
	****

FIGURE 10

10/16

		10	20	30	40	50	60
SeqID186	MASLGLQLVGYILGLLGLLGTLVAMLLPSWKTSSYVGASIVTAVGFSKGLWMECATHSTG						
	.....						
AF072128	MASLGVQLVGYILGLLGLLGTSIAMLLPNWRTSSYVGASIVTAVGFSKGLWMECATHSTG						
		10	20	30	40	50	60
		70	80	90	100	110	120
SeqID186	ITQCDIYSTLLGLPADIQAAQAMMVTSSAIISSLACIIISVVGMRCTVFCQESRAKDRVAVA						
	.....						
AF072128	ITQCDIYSTLLGLPADIQAAQAMMVTSSAMSSLACIIISVVGMRCTVFCQDSRAKDRVAVV						
		70	80	90	100	110	120
		130	140	150	160	170	180
SeqID186	GGVFFILGGLLGFIPVAWNLHGILRDFYSPLVPDSMKFEIGEALYLGIISSLFSLIAGII						
	.....						
AF072128	GGVFFILGGILGFIPVAWNLHGILRDFYSPLVPDSMKFEIGEALYLGIIISALFSLVAGVI						
		130	140	150	160	170	180
		190	200	210	220	230	
SeqID186	LCFSCSSQRNRSNYDAYQAQPLATRSSPRPGQPPKVKSEFNSYSLSLTGYV						
	.....						
AF072128	LCFSCSPQGNRTNYDGYQAQPLATRSSPRSAQPKAKSEFNSYSLSLTGYV						
		190	200	210	220	230	

FIGURE 11

98.3% identity in 121 aa overlap

```

                                10      20      30
seqID213                      RFRKETDNAAIIMKVDRQMVVLEEEFRNISPEELKME
                                .....
AB001993 MSDSLVVCEVDPTELTEKLRLKFRFRKETDNAAIIMKVDRQMVVLEEEFRNISPEELKME
                                10      20      30      40      50      60

                                40      50      60      70      80      90
seqID213 LPERQPRFVVYSYKYVRDDGRVSYPLCFIFSSPVGCKPEQQMMYAGSKNRLVQTAE LTKV
                                .....
AB001993 LPERQPRFVVYSYKYVHDDGRVSYPLCFIFSSPVGCKPEQQMMYAGSKNRLVQTAE LTKV
                                70      80      90      100     110     120

                                100     110     120
seqID213 FEIRTTDDLTEAWLQEKL SFFR
                                .....
AB001993 FEIRTTDDLTEAWLQEKL SFFR
                                130     140
```

FIGURE 12

95.6% identity in 91 aa overlap

```
seq ID191                                10      20
                                         MGCVFQSTEDKCIFKIDWTLS
W36955  MFCPLKLILLPVLLDYSGLNDLNVSPPELTVHVGDSALMGCVFQSTEDKCIFKIDWTLS
                                         10      20      30      40      50      60
seq ID191  30      40      50      60      70      80
PGEHAKDEYVLYYYSNLSVPIGRFQNRVHLMGDILCNDGSLLLQDVQEADQGTYICEIRL
W36955  PGEHAKDEYVLYYYSNLSVPIGRFQNRVHLMGDNLNDGSLLLQDVQDVE
                                         70      80      90     100     110
seq ID191  90     100
KGESQVFKKAVVLHVLPEEPKGTQMLT
```

FIGURE 13

99.0% identity in 381 aa overlap;

```

seqID200  10      20      30      40      50      60
          MLLSIGMLMLSATQVYTVLTVQLFAFLNPLPVEADILAYNFENASQTFDDLPARFGYRLP
          :
AF037204  10      20      30      40      50      60
          MLLSIGMLMLSATQVYTVLTVQLFAFLNLLPVEADILAYNFENASQTFDDLPARFGYRLP

id200     70      80      90      100     110     120
          AEGLKGFLINSKPENACEPIVPPVKDNSSGTFIVLIRRLDCNFDIKVLNAQRAGYKAAI
          :
AF037204  70      80      90      100     110     120
          AEGLKGFLINSKPENACEPIVPPVKDNSSGTFIVLIRRLDCNFDIKVLNAQRAGYKAAI

id200     130     140     150     160     170     180
          VHNVDSDDLISMGSDIEVLKKIDIPSVFIGESSASSLKDEFTYEKGHLLVPEFSLPL
          :
AF037204  130     140     150     160     170     180
          VHNVDSDDLISMGSDIEVLKKIDIPSVFIGESSANSLSLKDEFTYEKGHLLVPEFSLPL

id200     190     200     210     220     230     240
          EYYLIPFLIIVGICLILIVIFMITKFVQDRHRARRNRLRKDQLKKLPVHKFKKGDEYDVC
          :
AF037204  190     200     210     220     230     240
          EYYLIPFLIIVGICLILIVIFMITKFVQDRHRARRNRLRKDQLKKLPVHKFKKGDEYDVC

id200     250     260     270     280     290     300
          AICLDEYEDGDKLRILPCSHAYHCKCVDPWLTKTKKTCCPVCRQKVVP SQGDS S DTDSSQ
          :
AF037204  250     260     270     280     290     300
          AICLDEYEDGDKLRILPCSHAYHCKCVDPWLTKTKKTCCPVCRQKVVP SQGDS S DTDSSQ

id200     310     320     330     340     350     360
          EENEVTEHTPLLRLPLASVSAQSFGALSESRS HQNMTESSDY EEDDNEDTDSSDAENEINE
          :
AF037204  310     320     330     340     350     360
          EENEVTEHTPLLRLPLASVSAQSFGALSESRS HQNMTESSDY EEDDNEDTDSSDAENEINE

id200     370     380
          HDVVVQLQPNGERDYNIA NTV
          :
AF037204  370     380
          HDVVVQLQPNGERDYNIA NTV

```

FIGURE 14

100.0% identity in 68 aa overlap

	10	20	30	40	50	60
seqID192	MSVFWGFGVGLVPWFIPKGPNRGVIIITMLVTCSVCCYLFWLIAILAQLNPLFGPQLKNET					
	::					
Y15286	MSVFWGFGVGLVPWFIPKGPNRGVIIITMLVTCSVCCYLFWLIAILAQLNPLFGPQLKNET					
	20	30	40	50	60	70

seqID192	IWYLKYHW
	::::::::
Y15286	IWYLKYHW
	80

FIGURE 15

100.0% identity in 68 aa overlap

```

seqID201      -MDSRVS--SPEKQDKENFVGVNKRLGVCGWILFSLSFLLVIITFPISIWMLKIIREY
seqID227      -----MWLDP-----VFPLFPVG-----DH
X85116        MAEKRHTRDSEAQRLPDSFKDSPSKGLGPCGWILFAFSFLFTVITFPISIWMCIKIIEY
                *                **

seqID201      ERAVVFRLGRIQADKAKGPGILVLVPCIDVFVKVDLRTVTCNIPPQEILTRDSVTTQVDG
seqID227      Y-----LPHLHMDVLEG--LILVLVPCIDVFVKVDLRTVTCNIPPQEILTRDSVTTQVDG
X85116        ERAIIFRLGRILQGGAKGPGILFFILPCTDSFIKVDMTISFDPPEILTKDSVTISVDG
                *  .      *  *  .*** *  .***.***.  .*****.***  ***

seqID201      VVYYRIYSAVSANVNDVHQATFLLAQTTLRNVLGTQTLSQLAGREEIAHSIQTLDD
seqID227      VVYYRIYSAVSANVNDVHQATFLLAQTTLRNVLGTQTLSQLAGREEIAHSIQTLDD
X85116        VVYYRVQONATLAVANITNADSATRLLAQTTLRNVLGTKNLSQLSDREEIAHNMOSTLDD
                *****  *  *****  .** *****.*****.*****.  .***

seqID201      ATELWGIRVARVEIKDVRIPVQLQRSMAAEAEATREARAKVLAAEGEMSASKSLKSASMV
seqID227      ATELWGIRVARVEIKDVRIPVQLQRSMAAEAEATREARAKVLAAEGEMNASKSLKSASMV
X85116        ATDAWGIKVERVEIKDVKLPVQLQRMAAEAEASREARAKVIAAEGEMNASRALKEASMV
                **  .  ***.*** *****.*****.*****.*****.*****  **  .**  ****

seqID201      LAESPIALQLRYLQTLSTVATEKNSTIVFPLPMNILEGIGGVSYDNHKKLPNKA
seqID227      LAESPIALQLRYLQTLSTVATEKNSTIVFPLPMNILEGIGGVSYDNHKKLPNKA
X85116        ITESPAALQLRYLQTLTTIAAEKNSTIVFPLPIDMLQGIIGAKHSHLG-----
                .*** *****.*** *****.*****.*****.*****

```

FIGURE 16